Special Master's Report on the Scientific Foundations of STRmix™

Submitted to Magistrate Judge David T. Schultz

Unites States District Court,

District of Minnesota

In *United States of America v. Kenneth Davon Lewis*, No. 18-CR-194 (ADM/DTS)

October 31, 2019

Table of Contents

I. Introduction

The defendant in this case is charged with possession of a firearm as a felon. The government seeks to present expert testimony that DNA consistent with that of the defendant was found on a Smith & Wesson 9 mm pistol. DNA testing was performed at the Midwest Regional Forensic Laboratory, which used cotton swabs to collect human DNA from three areas of the pistol: the grip and trigger areas (Item 8-A), the slide serrations, safety and slide release levers and hammer (Item 8-B), and the empty firearm magazine (Item 10-A). The laboratory interpreted DNA profiles found on each of these items as a mixture of human DNA originating from four individuals.

In order to assess whether the defendant could have been a contributor to the DNA mixtures found on the pistol, the laboratory used a computer program called STRmix™ to analyze the mixed DNA profiles and to compare them to the defendant's profile. Based on the computer analysis, the laboratory reported that each of the three mixtures "is greater than one billion times more likely if it originated from the defendant and three unknown unrelated individuals than if it originated from four unknown unrelated individuals." (Report on the Examination of Physical Evidence, June 16, 2019; Gov't Exhibit 1). The government seeks to present expert testimony at trial in order to describe these findings and, presumably, explain their meaning. The defendant objects to admitting this evidence on grounds that "STRmix has not been sufficiently validated for analysis of complex DNA mixtures like those here" and hence

fails to meet standards for the admissibility of expert evidence set forth in *Daubert v. Merrell Dow Pharmaceuticals, Inc*., 509 U.S. 579 (1993) and Rule 702 of the Federal Rules of Evidence.[1]

The Court held an evidentiary hearing on the admissibility of the proffered expert testimony. Due to the complexity of the scientific issues, a Special Master was appointed to advise the court. The Special Master was charged with reviewing transcripts and exhibits from evidentiary hearings held on March 5 and April 2, 2019, attending and reviewing exhibits at an evidentiary hearing on August 16, 2019, and submitting this report to the court addressing the reliability of the principles and methods at issue in this case and whether they were reliably applied to the facts of the case.

In *Daubert*, the U.S. Supreme Court used the term "reliable" to refer to "*evidentiary* reliability–that is, trustworthiness."[2] It explained that: "In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity.*"[3] Accordingly, this report focuses on whether the scientific validity of STRmix™ has been established and whether the program produces trustworthy results. This requires consideration of two related issues: (1) whether STRmix™ is a "reliable method" and (2) whether the government's experts have "reliably applied" STRmix™ in this case.

This report will begin by discussing recent scientific commentary on the scientific validity of STRmix™, focusing particularly on an authoritative 2016 report of the President's Council of

---

[1] Rule 702 states: A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
(a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
(b) the testimony is based on sufficient facts or data;
(c) the testimony is the product of reliable principles and methods; and
(d) the expert has reliably applied the principles and methods to the facts of the case.

[2] 509 U.S. at 579 n. 9

[3] *Id.* (emphasis in original).

Advisors on Science and Technology (PCAST, 2016; Defense Exhibit 2). It will then provide an overview of the positions taken by the parties in this case regarding the scientific status of STRmix™ and whether it is ready to be used in court. In order to evaluate the positions of the parties, it will then be necessary to delve more deeply into the scientific details of how STRmix™ works and how it might possibly fail. The scientific discussion will provide a basis for the final section of this report, which analyzes the strengths and weaknesses of the evidence presented.

II. The PCAST Report

A useful starting point for an analysis of the scientific validity of STRmix™ is the 2016 PCAST report entitled "Forensic Science in Criminal Courts: Ensuring the Validity of Feature-Comparison Methods."[4] PCAST has been described as "an advisory group of the nation's leading scientists and engineers who directly advise the President and the Executive Office of the President."[5] In 2015 President Obama asked PCAST to evaluate ways to strengthen forensic science in the United States. In 2016, PCAST issued a report that commented on the steps needed to assure the scientific validity of "feature-comparison methods" in forensic science; the report also provided evaluations of the scientific validity of methods used in six forensic science disciplines that rely on "feature comparison." One of those disciplines was "DNA analysis of complex-mixture samples" (PCAST, p. 75-83). The report makes it clear that DNA mixtures like the ones analyzed in this case are considered to be complex-mixture samples.

[4] The parties appeared to agree with PCAST's framing and analysis of the scientific issues. They differ mainly in their assessment of whether research published *after* the release of the PCAST report is sufficient to address issues and limitations noted in the PCAST report. Consequently, a detailed discussion of the PCAST report is a useful introduction to this scientific debate.

[5] See, https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/about

PCAST noted that forensic DNA laboratories had initially relied on subjective judgment of examiners, along with a simplified statistical procedure known as CPI (cumulative probability of inclusion), to interpret complex DNA mixtures.[6] PCAST found that "DNA analysis of complex mixtures—defined as mixtures with more than two contributors—is inherently difficult and even more for small amounts of DNA" (PCAST p. 75). It cited research showing that interpretations of complex DNA mixtures often differ greatly among different examiners and different laboratories. It cited differing interpretations emerging from laboratories in Texas as evidence that "the problems with subjective analysis of complex DNA mixtures were not limited to a few individual cases: they were systemic." (p. 77). PCAST concluded that "the interpretation of complex DNA mixtures with the CPI statistic has been an inadequately specified—and thus inappropriately subjective—method [that is] clearly not foundationally valid." (p. 78). It discussed efforts by the Texas Forensic Science Commission to produce a scientific article, published in 2016,[7] that sets clearer rules for interpreting complex mixtures using CPI statistics, but noted that PCAST "has not had adequate time to assess whether the rules are … sufficient to define an objective and scientifically valid method for the application of CPI" (PCAST, p. 78).

STRmix™ is one of several computer programs that have been developed in order to provide a more objective method for the analysis of complex DNA mixtures. These programs use mathematical algorithms to perform *probabilistic genotyping (PG)*, which involves a

[6] At various points in the hearing transcript witnesses refer to CPE (cumulative probability of exclusion), which is the same technique as CPI. The technique attempts to determine the proportion of a population that could be included (CPI) or excluded (CPE) as a possible contributor to a mixture. The two terms are complementary, i.e., CPI = 1 – CPE.

[7] Bieber, Buckleton, Budowle, Butler & Coble (2016).

systematic assessment how likely the observed mixture is to occur under various assumptions about the possible contributors. This assessment is then used to evaluate the strength of the DNA evidence for proving that a given individual contributed (or did not contribute) DNA to a mixed sample. According to the PCAST report:

> These probabilistic genotyping software programs clearly represent a major improvement over purely subjective interpretation. However, they still require careful scrutiny to determine (1) whether the methods are scientifically valid, including defining the limitations on their reliability (that is, the circumstances in which they may yield unreliable results) and (2) whether the software correctly implements the methods. (PCAST p. 79).

PCAST reviewed a number of studies designed to test the validity of STRmix™ and a similar program known as TrueAllele™. Its conclusion (in September 2016, when the report was issued) was as follows:

> Most importantly, current studies have adequately explored only a limited range of mixture types (with respect to number of contributors, ratio of minor contributors, and total amount of DNA). The two most widely used methods (STRmix and TrueAllele) appear to be reliable within a certain range, based on the available evidence and the inherent difficulty of the problem. Specifically, these methods appear to be reliable for three-person mixtures in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method. For more complex mixtures (e.g. more contributors or lower

proportions), there is relatively little published evidence. (PCAST p. 80-81)(footnotes omitted).

In January 2017, PCAST issued an addendum to its report that included additional discussion of probabilistic genotyping systems in light of comments received on the initial PCAST report. The addendum reiterated PCAST's discussion of the difficulty of interpreting complex DNA mixtures, noting that:

> Early efforts to interpret these profiles involved purely subjective and poorly defined methods, which were not subjected to empirical validation. Efforts then shifted to a quantitative method called combined probability of inclusion (CPI); however, this approach also proved seriously problematic. (PCAST, 2017, p. 8)

The addendum also discussed a controversy that arose in a murder case in New York in which STRmix™ and TrueAllele™ "reached opposite conclusions about whether a DNA sample in the case contained a tiny contribution (~1%) from the defendant." PCAST convened a meeting with the developers of the two programs (John Buckleton for STRmix™ and Mark Perlin for TrueAllele™) to discuss "how best to establish the range in which a PG software program can be considered to be valid and reliable." (PCAST, 2017, p. 8). After considering presentations by both developers, PCAST rejected Perlin's contention that it was "mathematically impossible for the likelihood ratio approach in his software to incorrectly implicate an individual" (p. 8). PCAST instead endorsed Buckleton's view that empirical testing of the software with different kinds of mixtures is necessary in order to determine how well it works and to establish the conditions under which it will and will not produce trustworthy results. The addendum's commentary on probabilistic genotyping programs concluded with the following suggestions:

The path forward is straightforward. The validity of specific PG software should be validated by testing a diverse collection of samples within well-defined ranges. The DNA analysis field contains excellent scientists who are capable of defining, executing, and analyzing such empirical studies.

When considering the admissibility of testimony about complex mixtures (or complex samples), judges should ascertain whether the published validation studies adequately address the nature of the sample being analyzed (e.g., DNA quantity and quality, number of contributors, and mixture proportion for the person of interest). (PCAST, 2017, p. 9)

III. Positions of the Parties

The parties in this case appear to agree with PCAST's conclusion that validation of STRmix™ requires empirical studies to test the performance of the program. Neither party offered the argument (attributed to Dr. Mark Perlin) that the mathematical algorithms used in the program are incapable of producing wrong or misleading conclusions. The major difference between the parties was their position on whether research performed to date is sufficient to establish that STRmix™ meets the requirements of evidentiary reliability under *Daubert* and Rule 702.

The government relies heavily on two validation studies published after the PCAST report that test the performance of STRmix™ at distinguishing contributors and non-

contributors to DNA mixtures (Bright, Richards, Kruijver, et al., 2018[8]; Moretti, Just, Kehl, et al. 2017). The study by Bright et al. assessed the performance of STRmix™ when analyzing 2825 DNA mixtures that had been prepared and analyzed at 31 different forensic DNA laboratories. The mixtures contained DNA of between three and six known individuals that had been mixed together in varying proportions. According to the authors the study "was done in accordance with the specific manner outlined in the PCAST report" and "demonstrates a foundational validity [of STRmix™] for complex, mixed DNA profiles to levels well beyond the complexity and contribution levels suggested by PCAST." (Bright et al., 2018, p. 23). While there were a few instances in which STRmix™ produced results that falsely linked non-contributors to the mixtures, these misleading results were rare and occurred no more often than would be expected by chance due to adventitious (coincidental) similarity between DNA profiles of different individuals. In other words, the rate of false inclusions was approximately what would be expected if STRmix™ performed its function flawlessly. The rate of false exclusions was somewhat higher, but low enough to support claims that the program is highly accurate.[9]

The study by Moretti et al. assessed the performance of STRmix™ when analyzing 277 DNA mixtures prepared at the FBI laboratory. These mixtures contained DNA of between two and five known individuals that had been mixed together in varying proportions (84 were four-person mixtures; 24 were five-person mixtures). According to Moretti et al., the results showed overall that STRmix™ software performed as expected. "With very few exceptions, genotype weights were intuitively correct, and the statistical results were consistent with scientific expectations" (Moretti et al., 2017, p. 143). Moretti et al. concluded that the results "establish

[8] Government Exhibit 16; Defense Exhibit 3).
[9] The error rate in this study is discussed in more detail later in this report, in Section V.*A.2.*

that STRmix…is fit for purpose for the interpretation and statistical assessment of single course profiles and mixtures originating from two, three, four, and five individuals." (Moretti et al., 2017, p. 143).

The government's experts, Dr. Buckleton and Ms. Anne Ciecko, the Technical Leader of the DNA unit of the Midwest Regional Forensic Laboratory, both took the position that these studies are more than sufficient to establish that STRmix™ is a scientifically valid method that produces trustworthy results. The defendant's experts, by contrast, suggested that the two post-PCAST validation studies, while valuable for understanding the performance of STRmix™, do not go far enough in testing its performance.

The defense offered several criticisms.

A. Lack of Independent Validation

The PCAST report emphasized that the validation of probabilistic genotyping software should be carried out by independent groups or organizations, and not just by the developers of the software.

> Appropriate evaluation of the proposed methods should consist of studies by multiple groups, *not associated with the software developers*, that investigate the performance and define the limitations of programs by testing them on a wide range of mixtures with different properties. (PCAST, p. 79)(emphasis in original).
>
> [S]uch studies should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome (PCAST, p. 81).

In this case the defense suggested that the FBI laboratory, and the 31 laboratories that collaborated on the Bright et al. study, have a stake in the outcome of the research because they had all previously made the commitment to purchase, validate, and train staff to use STRmix™. Most of these laboratories had already begun using STRmix™ in casework. The defense suggested that these labs had already made a costly commitment to use this software and hence had an incentive to present its performance in the best light.

With regard to the Bright et al. study, the defense pointed out that all of the statistical analysis and data interpretation underlying the article, as well as the framing of conclusions, was carried out by the developers of STRmix™. The data analysis and interpretation was carried out by ESR, which is a "Crown Research Institute" operated by the government of New Zealand. ESR is the organization that markets STRmix™ as a commercial product. One of the principle developers of the software, Dr. John Buckleton, who testified for the government in this case, is an employee of ESR. Dr. Buckleton testified that he wrote "large parts" of the article by Bright et al. and was deeply involved with the conception and execution of the research project described in the article (TR p. 91).

B. Failure to Establish Boundaries of Validity

The defense experts argued that there are two aspects to validation of a software product like STRmix™. One aspect is a demonstration that the product works as expected and can produce accurate results under specified conditions. The second aspect is an exploration of circumstances under which the software will not work as intended in order to establish the boundaries beyond which it will not produce trustworthy results. In his testimony, Prof. Dan Krane, a biology professor from Wright State University, referred to these two aspects as the

"yin and yang" of validation. The defense experts suggested that the post-PCAST validation studies satisfy the first condition but not the second. These studies show that the software can work very well (under the particular circumstances used in the studies), but provide insufficient information to assess when it might not work well. Prof. Krane suggested that STRmix™ may be particularly vulnerable to failure in instances where there is uncertainty about the number of contributors, where there are large differences in the proportion of DNA from each contributor, and where the DNA test may have failed to detect the full DNA profiles of all contributors.

C. Failure to Comply with the Highest Standards for Software Verification and Validation

During the hearing there was much discussion of standards for software validation. Dr. Buckleton claimed that STRmix™ has been validated in a manner consistent with standards or guidelines issued by several organizations, including the Scientific Working Group for DNA Analysis Methods (SWGDAM), which is sponsored by the FBI (Defense Exhibit 1), the International Society for Forensic Genetics (ISFG)(Defense Exhibit 21), which is based largely in Europe, and the Forensic Science Regulator (FSR) of Great Britain (Defense Exhibit 22). All three of these bodies have issued guidelines that speak specifically to the validation of probabilistic genotyping software.

The defense experts asserted that STRmix™ falls short of compliance with the SWGDAM, ISFG and FSR guidelines, but their major objection was that STRmix™ should be evaluated under higher standards promulgated by the Institute of Electronic and Electrical Engineers (IEEE) for "safety critical systems." (Defense Exhibit 20). The validation requirements imposed by the IEEE standards vary in stringency depending on the consequences of a software failure. When

the consequences of a software failure would be very serious (e.g., loss of life, major financial or social loss) the standards require that the validation be performed in a manner that is independent of software development, and specifically that it be done by people who were not part of the development team and who have financial and managerial independence from the developers.

Dr. Buckleton testified that he thought it "perfectly fair" to suggest that STRmix™ should be evaluated according to the highest IEEE standards (TR p. 108, Ln 11) given the critical role that STRmix™ may play in criminal prosecutions. He argued that the existing validation "meets or very nearly meets" IEEE's requirement of independence. The defense experts disagreed. The primary defense expert on this issue was Prof. Mats Heimdahl, head of the Computer Science and Engineering Center at the University of Minnesota, who specializes in software engineering for critical safety systems. He reviewed various documents provided by ESR on the "verification and validation" of STRmix™ and found ESR's efforts insufficient for "a safety critical or life critical system." (TR p. 575, Ln. 22).[10]

Among the deficiencies Prof. Heimdahl cited were: (1) lack of a "hazard analysis" to identify possible failure points in the software for special scrutiny; (2) lack of "formal specifications" of what the software is expected to do that can provide a basis for independent assessment of whether it is doing what is expected; (3) insufficient documentation of changes in the software which interferes with the "traceability" (i.e., the effort to link performance characteristics of the software to particular features of the code); and (4) the lack of a formal

[10] Prof. Heimdahl did not, however, take a position on whether STRmix™ should be considered a safety critical system, saying that would depend on how the software is used in practice.

"code inspection" by an independent party to test and confirm that the software is operating according to specifications.

In explaining why he considers it important that evaluation of the software be independent of development, Prof. Heimdahl echoed Prof. Krane's concerns about the failure to establish the boundaries of validity:

> If you're developing the software, it's your baby. Subconsciously or consciously, people don't want to break it, and there's data supporting this. If you have [an independent] testing team, their job is not to show that this software works. That's somebody else's job. Their job is to break it. (TR 583, LN 11-16)

According to Prof. Heimdahl, complex software often proves to be "brittle," which means it can work well under some conditions while being vulnerable to unexpected catastrophic failure in other conditions (TR p. 593, Ln 1). Consequently, when validating safety critical software it is vital to explore possible points of failure by trying to "break it," and thereby expose any flaws and limitations before the software is routinely used.

The third defense witness was Nathan Adams, a software engineer who works for a consulting company operated by Professor Krane. Mr. Adams emphasized the importance of "verification and validation" (V&V) of software that is used in critical systems. He argued that careful V&V is particularly important for STRmix™ because it is difficult to assess the correctness of the program's statistical outputs. For example, there is no way to assess whether the statistical estimates provided in this case (that the observed mixtures are *one billion times* more probable if the defendant was a contributor) are precisely correct. While the validation studies show that STRmix™ is good at distinguishing contributors from non-

contributors, these studies do not prove that the statistics the program generates to describe the strength of this evidence are always accurate. We can have confidence in these statistics, Adams argued, only if we know that the program is computing the statistics in exact accordance with accepted modeling methods. The problem with STRmix™, according to Mr. Adams, is that the program has not been documented sufficiently to allow independent experts to assess whether it is working as intended. He agreed with Prof. Heimdahl that the program lacked the "formal specifications" necessary for a successful independent review. Adams based this conclusion on his own review of proprietary information about the program and its documentation.[11]

IV. Scientific and Technical Background on Forensic STR Analysis and STRmix

At this juncture it will be helpful to delve more deeply into the details of how DNA profiles are produced and how STRmix™ interprets DNA profiles. This technical information provides a framework and terminology for discussing the strengths and weaknesses of the evidence and arguments offered by the parties. This information is necessary for assessing the significance of the concerns raised by the defense experts with respect to the legal issues the court must decide.

[11] Mr. Adams testified that ESR had required him to sign a non-disclosure agreement (NDA) when he first reviewed the source code for STRmix and related documentation in another case. He stated that the NDA prevented him from discussing some of the observations he had made at that time that might be relevant in this case. The defense offered in evidence a letter from a law firm representing ESR that threatened to seek legal remedies against Adams if he testified in a manner that violated the NDA (Defendant's Exhibit 19).

This section will first discuss the biological side of DNA analysis, focusing on how a forensic laboratory can determine the genetic characteristics of biological samples. It will then discuss the statistical side of the analysis, focusing on how STRmix™ assesses whether a particular individual contributed DNA to a mixed sample of biological material.

A. The Biological Analysis: Generating DNA Profiles[12]

The PCAST report provided a concise description of the process used to generate a DNA profile:

> … DNA is first chemically *extracted* from a sample containing biological material, such as blood, semen, hair, or skin cells. Next, a predetermined set of DNA segments ("loci") containing small repeated sequences are *amplified* using the Polymerase Chain Reaction (PCR), an enzymatic process that replicates a targeted DNA segment over and over to yield millions of copies. After amplification, the lengths of the resulting DNA fragments are *measured* using a technique called capillary electrophoresis, which is based on the fact that longer fragments move more slowly than shorter fragments through a polymer solution. The raw data collected from this process are analyzed by a software program to produce a graphical image (an electropherogram) and a list of numbers (the DNA profile) corresponding to the sizes of the each of fragments (by comparing them to known "molecular size standards")…. (PCAST, 2016, p. 69)(footnotes omitted).

A close look at the DNA testing conducted in this case provides an illustration. At the Midwest Regional Forensic Laboratory, analysts used cotton swabs to collect cellular material from various parts of the Smith & Wesson pistol. They *extracted* the DNA by using chemicals to

[12] Butler (2012) is a general reference that supports and provide additional background on the information presented in this section.

break open the cells and release the DNA, and additional chemicals to clean and purify the DNA for analysis. Next they *amplified* DNA found at 24 *loci* using a commercial test kit called the Promega PowerPlex Fusion System and a laboratory instrument known as a thermal cycler (which used *PCR* to create millions of copies of DNA fragments from the targeted loci). The analysts then injected the DNA samples into a computer-controlled instrument called the Applied Biosystems 3500 Genetic Analyzer, which used *capillary electrophoresis* to measure the length of the amplified DNA fragments. The Genetic Analyzer then produced *electropherograms* showing the lengths of the DNA fragments that were detected at each locus of each sample.

The DNA fragments examined in this process originate at loci that contain short, repeating sequences of genetic code called short tandem repeats (STRs). The number of repetitions tends to vary from person to person, which causes the length of these DNA fragments also to vary in ways that can be used to distinguish different individuals. At each locus there are several possible lengths that the fragments might have. Each possible length variant is called an *allele*.[13] The alleles are identified by numbers that correspond to the number of "repeats" in the STR. For example, a fragment containing eleven repetitions of a short sequence of genetic code will be labeled "allele 11."

At each locus a person generally inherits two of these alleles, one from each parent. For a given individual, the same pair of alleles will be found at that locus in the DNA of all their nucleated cells and remains consistent throughout their lifetime. Different individuals tend to

[13] The Genetic Analyzer identifies the alleles by measuring the lengths of the DNA fragments (using a process known as capillary electrophoresis). It is the results of this analysis that are displayed in graphs known as electropherograms, such as the one shown below (Figure 1).

have different alleles. While two people may, by chance, have the same alleles at a few loci, the chances of such a coincidence diminish rapidly as more loci are examined. The set of alleles that a person has across multiple loci is called a multi-locus genotype or a DNA profile.

Figure 1 displays one of several electropherogram produced in this case for Item 8-A, the swab from the grip and trigger area of the pistol. It shows the alleles detected at five genetic loci. The names of the five loci are D8S1179, D12S391, D19S433, FGA, and D22S1045. Other electropherograms (not shown here) display the alleles detected at 19 other loci when Item 8-A was tested.

The electropherograms display a set of peaks that signal the presence of alleles. The position of a peak along the graph indicates the length of the amplified DNA fragments containing a STR. Based on the position of the peak (relative to molecular size standards which are not shown here), the computer determines which allele the peak represents and applies a label shown at the top of the box seen immediately below the peak on the electropherogram. The first set of peaks, on the left side of the electropherogram, show the alleles detected at locus D8S1179. There are seven peaks (indicating seven alleles) which the computer has determined to be alleles 9, 10, 11, 12, 13, 14 and 15. By contrast, at the far right side of the electropherogram are two peaks detected at locus D22S1045, which the computer has determined to be alleles 12 and 15.

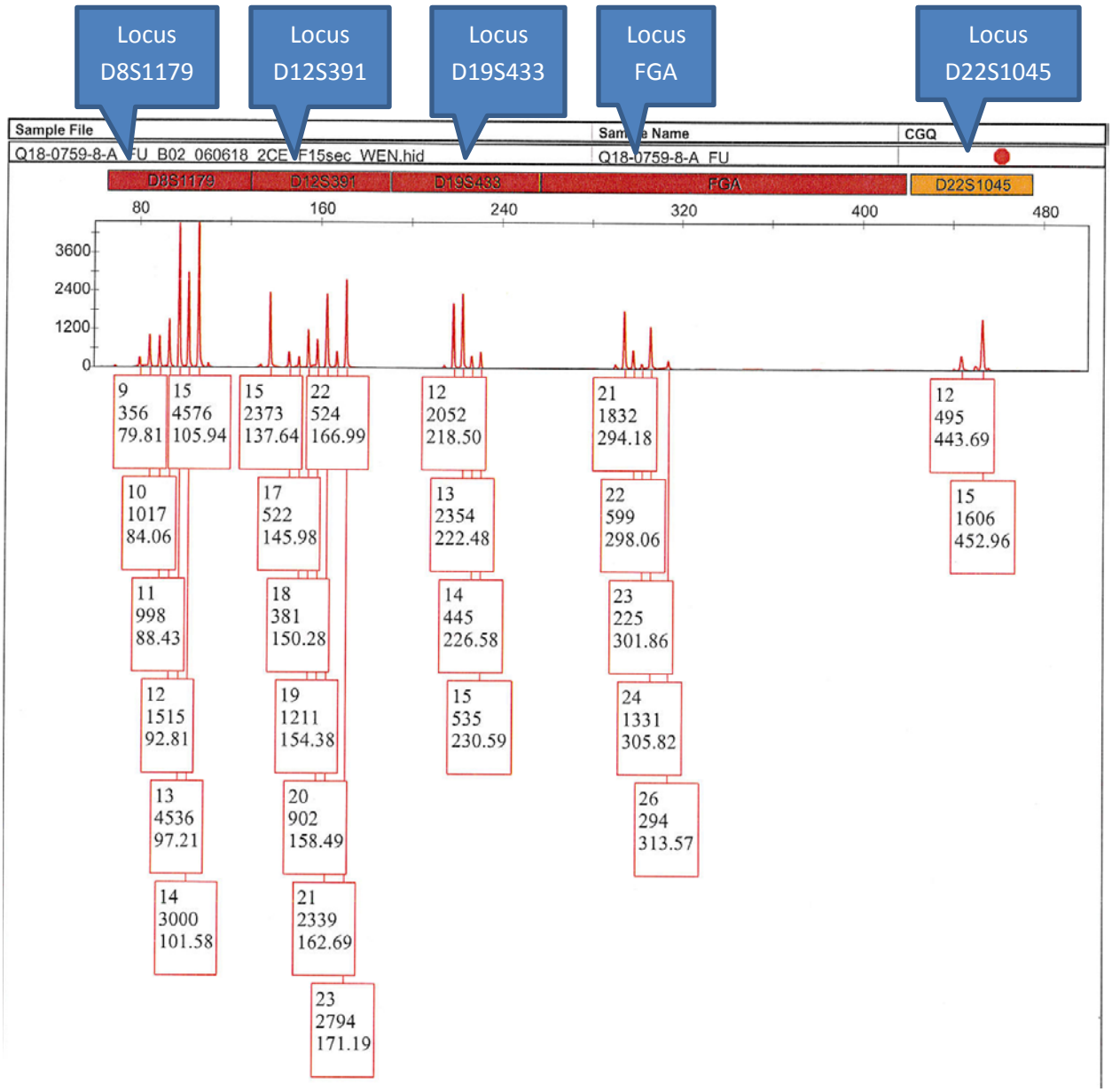

Figure 1: Electropherogram Showing Alleles Detected at Five Genetic Loci (From Left to Right: D8S1179, D12S391, D19S433, FGA, and D22S1045). Labels under the peaks indicate (from top to bottom) the allele detected, the height of the peak (in RFU), and the peak's measured position on the x-axis.

The height of each peak corresponds roughly to the quantity of DNA detected—taller peaks indicate more DNA. The second number in the box under each peak indicates the measured height of the peak in "relative fluorescent units" (RFU). The third number at the bottom of each box is a measurement (not important for this discussion) indicating where the peak falls along the X-axis of the electropherogram.

From this electropherogram, a DNA analyst can draw inferences about DNA Sample 8-A. First, the large number of alleles indicates the sample contains DNA of more than one individual. Because each individual generally contributes at most two distinct alleles at each locus, the presence of seven alleles at locus D8S1179, and eight alleles at locus D12S391, indicates there are at least four contributors to this DNA mixture. Second, notice that there is considerable variation in the height of the peaks that represent these alleles at locus D8S1179 and D12S391. If we assume four contributors, this height variation indicates that they are contributing different amounts of DNA to the mixture.

There is considerable uncertainty, however, about the genotypes of individual contributors to this mixture. In general, the two alleles of each contributor will produce peaks of roughly equal height. So an analyst might infer that the two tallest peaks found at locus D8S1179, alleles 13 and 15, came from a single individual who contributed more DNA to the mixture than the other contributors, but that is not the only possibility. The peaks representing alleles 13 and 15 might also be taller because multiple individuals share those alleles. For example, a mixture of DNA from three individuals who have genotypes 10,13; 11,13; and 12,13 would produce a tall 13 peak and shorter peaks at 10, 11 and 12 (as seen in this electropherogram). So possibilities of that nature must also be considered. Furthermore,

individuals sometimes inherit the same allele from both parents. When DNA of these homozygous individuals appears in the mixture it produces a single peak of roughly double the height of a heterozygous peak. So it is also possible that alleles 13 and 15 each came from a different homozygous individual. The number of combinations of possible genotypes that might account for the peaks observed at a given locus will vary depending on the number of peaks and the analyst's assumptions about the number of contributors, but it could easily number in the hundreds or even thousands.

Assessment of the genotypes of possible contributors is further complicated by the possibility that the test may have failed to detect all alleles of all contributors. Notice, for example, that only two alleles were detected at locus D22S1045. That might have occurred because the contributors to the mixture collectively have no alleles other than the two alleles that were detected (12 and 15). Another possibility, however, is that the test lacked the sensitivity to detect some of the contributors' alleles at this locus. A variety of factors affect the sensitivity of the test, including the degree to which the DNA has been degraded due to age, environmental exposure, or conditions of storage. There is variation among the loci in their overall sensitivity (i.e., in the amount of DNA needed to produce a detectable allele) and in their susceptibility to loss of signal due to degradation.

Another complicating factor is that peaks are sometimes produced spuriously due to the presence of small amounts of contaminating DNA (this is called "allelic drop-in") or due to technical limitations of the PCR process that produce small false peaks adjacent to large peaks due to a phenomenon known as "stutter." In mixed DNA samples it is often impossible to

distinguish false peaks produced due to allelic drop-in or "stutter" from true peaks originating from a minor contributor or a contributor whose DNA is somewhat degraded.

To evaluate various theories regarding the genotypes of contributors, the analyst must also take into account uncertainty about the number of contributors. For example, the theory that alleles 13 and 15 at locus D8S1179 each came from a homozygous contributor leaves five other peaks to be explained, which implies that there were three additional contributors (for a total of five contributors to the mixture). So the analyst will need to weigh the probability that there could be as many as five contributors when assessing the plausibility of this theory, which requires consideration of matters such as the number of alleles detected at other loci, taking into account, of course, uncertainty about the sensitivity of the test, whether all alleles were detected, and whether some of the peaks could be spurious.

This relatively brief discussion should help explain the difficulty analysts face when they are required to analyze mixed DNA profiles relying solely on subjective judgment. There are a multitude of factors to weigh and consider and these factors interact in complex ways. There are literally thousands of possible theories that might account for the peaks observed in a complex mixture, and all of these possible explanations must be considered and weighed against one another. A full evaluation of all possible explanations for the observed data will strain and may overwhelm the cognitive capacity even of a highly diligent and intelligent analyst. Hence it is not surprising that different analysts who attempt the interpretation may reach different conclusions about whether a particular individual might be a contributor and how strongly the individual is implicated by the DNA test results. In addition to being unreliable, analysts' interpretations may also be biased (PCAST, 2016, p. 76-77). If analysts are

exposed prematurely to information about the profiles of suspects, for example, it can cause an unconscious bias in favor of more incriminating interpretations of evidentiary DNA samples (deKeijser et al., 2016; Thompson, 2009).[14]

B. The Statistical Analysis: STRmix™

Promoters of STRmix™ claim it is more accurate than human analysts. Moreover, because it is an automated process, promoters claim that STRmix is not susceptible to the kinds of bias that may influence human analysts.

The central element of STRmix™ is a program that computes the probability that a mixture of DNA from a specified number of individuals, who have specific genotypes, will produce a set of peaks matching those observed in an electropherogram. In order to make this computation, the program needs information about the operating characteristics of the DNA test kit and genetic analysis software used by the laboratory. For example, it needs information on the variance of peak heights within individuals; the variance in heights of spurious peaks that are produced by *stutter* or *drop-in;* and how often, and under what circumstances, there is "drop-out" (failure to detect a peak). These values (called parameters) will vary from laboratory to laboratory depending on the details of how the lab has implemented their biological testing system. So each laboratory needs to perform a set of studies to establish the relevant parameters as part of the process of "internal validation." The parameter values (inferred from the studies) are fed into STRmix™ as "operator inputs." As Dr. Buckleton described it, these inputs allow the operator to "tune" STRmix™ to the operating characteristics

[14] This form of bias can be mitigated through the use of "sequential unmasking" procedures, which are designed to limit or delay exposing analysts to information that is unnecessary or that might be biasing if presented prematurely (*see*, Risinger, et al. 2002; Krane et al., 2008; Stoel et al, 2015; Dror et al., 2015; National Commission, 2015), but not all forensic DNA laboratories have adopted rigorous interpretation procedures of this type.

of the DNA profiling system of the particular laboratory in which it is used (TR p. 11). The operator also tells the program what to assume about the number of contributors to the mixture.

To analyze a mixed DNA profile like the one (partially) shown in Figure 1, STRmix™ first deduces all possible sets of genotypes that might explain the peaks observed in the electropherogram, relying on information provided by the laboratory about such matters as stutter percentages, the likelihood of drop-out and drop-in, and the operator's determination of the number of contributors. The program then uses a form of statistical modeling called Markoff Chain Monte Carlo (MCMC) to assign "weights" to possible genotype combinations based on how well each possible combination explains the observed peaks.

MCMC has been widely used in the field of statistics to model complex situations (Kruschke, 2011). The algorithm is an "iterative re-sampling process" that proposes and tests millions of combinations of genotypes and biological parameters that might possibly explain the observed peaks (Taylor, Bright and Buckleton, 2013). The biological parameters that the program proposes include the relative amount of DNA from each proposed contributor, the level of degradation of each contributor's DNA, and the amplification efficiency (reflected in relative peak heights) of each locus examined in the DNA profile. [15] The program begins by randomly proposing scenarios that might explain the peaks observed in the electropherogram. Each scenario is a guess regarding the genotypes of each contributor and the biological parameters. Each scenario is evaluated according to how well it explains the peaks observed in the electropherogram of the DNA mixture. Some scenarios are effectively ruled out because

[15] Details of the modeling process can be found in Taylor, Bright & Buckleton (2013).

they cannot explain the observed results (i.e., the probability of the observed results is at or near zero under the hypothesized scenario). Scenarios that cannot be ruled out are assessed according to how well they fit the observed results shown in the electrophergram.[16] In an iterative process, the computer compares the "fit" of each guess with the previous guess. This allows the computer gradually to improve its guesses in a manner analogous to the children's game of hot and cold. The algorithm generally favors guesses that produce a better fit to the peaks observed in the electropherogram (these are "hotter"); it makes these favored guesses more often, which causes the process to gravitate toward scenarios that provide better explanations for the observed peaks. In order to avoid prematurely focusing on a particular scenario, however, the program sometimes makes "colder" guesses in order to assure that all possible scenarios are considered.[17] Ultimately, the system evaluates the relative plausibility of various scenarios by the amount of time the program spends considering each scenario as it shuffles through various possibilities. The key output is a set of values, described as "genotype weights," that reflect the relative plausibility of various combinations of genotypes that the contributors might have at each locus. The output also includes the best-fitting estimates of each of the biological parameters.

In the final step of the analysis, STRmix™ uses the genotype weights and data on the frequency of the alleles in various reference populations to compute likelihood ratios (LRs),

[16] "Better fit" means there is a higher estimated probability of observing the peaks seen in the electropherogram. In order to compute the probability of the observed peaks under each scenario, the program relies on information established during the internal validation process concerning such matters as peak height variation, stutter percentages, drop-out and drop-in probabilities, and other operating characteristics of the biological analysis as performed by the instruments used in the laboratory in question.

[17] STRmix makes "hot" and "cold" guesses in accordance with an algorithm known as Metropolis-Hastings, which has proven to be a helpful method for assuring good performance by the MCMC process in assessing the likelihood of various scenarios (Taylor,Bright & Buckleton, 2013).

which are statements about the relative probability of the observed peaks in a mixed DNA sample under alternative hypotheses about who contributed DNA to the mixture. The mathematical formuli for computing LRs are described in Taylor, Bright and Buckleton (2013). In this case the analyst considered two alternative hypotheses about each of the DNA mixtures found on the pistol: (1) that the mixture consists of DNA of the defendant and three unrelated, unknown individuals; and (2) that the mixture consists of DNA of four unrelated, unknown individuals (but did not include DNA of the defendant).

Each LR that STRmix™computes is based, in part, on the frequency of alleles in a particular reference group (e.g., African-Americans; Caucasian-Americans; Asian-Americans; Hispanic-Americans). In this case the laboratory computed a separate LR under the assumption that the unknown contributors were members of each major reference group; all of these LR values exceeded a value of 1 billion. The laboratory chose not to distinguish among races when reporting the LRs; for each item the lab report simply said: "This mixture is greater than one billion times more likely if it originated from the defendant and three unknown unrelated individuals than if it originated from four unknown unrelated individuals."

STRmix™ is also capable of computing LR's for other pairs of alternative hypotheses. In a sexual assault case where the DNA of the complainant is found in an intimate sample, for example, the hypotheses to be compared might be: (1) that the mixture contains DNA of the complainant and a suspect; and (2) that the mixture contains the DNA of the complainant and an unrelated unknown person. If there were three contributors, the hypotheses might be: (1) that the mixture contains DNA of the complainant, the suspect and an unknown person; and (2) that the mixture contains DNA of the complainant and two unknown persons. By using data on

the frequency of the alleles in various reference populations, these LRs can be calculated under various assumptions about the race or ethnicity of the unknown persons.

LRs can also be used to assess hypotheses that assume an unknown contributor is a relative of the suspect (e.g., a brother, father, uncle or cousin), rather than an unrelated individual. These LRs are useful for evaluating the plausibility of the theory that DNA found in a mixture came from a person with a given degree of relatedness to the suspect (e.g., a brother) rather than the suspect himself.[18]

Because LR's are difficult to understand (Thompson, 2018), it is important to explain what they mean and what they do not mean. The LR is a statement about the relative probability of finding the peaks that were observed in a DNA mixture under two alternative hypotheses about the contributors to the mixture. LRs tell us how strongly the DNA evidence supports one hypothesis over the other, with higher numbers indicating a higher level of support.[19] But LRs do not tell us which hypothesis is true, nor do they indicate the probability that the favored hypothesis is true. In order to explain why this is so, a recent article in *Judicature* offered the following, extended illustration:

> Imagine that a bloodstain of recent origin is found at the scene of a crime. Imagine further that the DNA profile of the bloodstain is somehow determined to be the same as the DNA profile of rock-and-roll legend Elvis Presley. Finally, imagine that the DNA

[18] LRs calculated in this case would differ, and would likely be much lower, if one or more of those unknown persons was assumed to be a relative of the defendant. One of the helpful features of STRmix is that is allows rapid computation of LRs under a variety of different hypotheses regarding the origin of a DNA mixture. It is important for legal counsel to be aware of this capability and to use it to explore all hypotheses that might be relevant in a given case.

[19] Experts typically make the favored hypothesis the numerator of the LR so that LRs can range from one to infinity. A value of one means the evidence is equally probable under the two hypotheses, and hence that the evidence has no probative value for distinguishing between the hypotheses. A value greater than one means that the evidence is more likely under the favored hypothesis, and hence that the evidence supports the favored hypothesis.

> profile in question is one million times more likely to be observed if the sample came from Elvis than if it came from a random person. Based on the DNA evidence, what can the examiner logically infer about the probability that the crime scene stain came from Elvis Presley?
>
> A moment of reflection should be sufficient to realize that the examiner can draw no conclusion about the probability that the crime scene stain came from Elvis based on the DNA evidence alone; the examiner must also consider other matters, such as whether Elvis could plausibly be the source. In this case, the suspect (Elvis) has a strong alibi — he was widely reported to have died in 1977. If the forensic scientist believes this "alibi," then the probability that the bloodstain came from Elvis is necessarily zero. (Thompson, Vuille, Taroni & Biederman, 2018, p. 25).

In this illustration, the LR of one million provides valuable information. It tells us that the DNA profile found in the bloodstain is very uncommon if it came from someone other than Elvis, which means the DNA evidence is strongly incriminating for Elvis. But the LR does not tell us how likely Elvis is to be the source of the bloodstain. To draw conclusions about that question, one would need to consider all the other evidence bearing on whether Elvis could be the source. And the same will be true of every defendant incriminated by DNA evidence. By itself, the DNA evidence can never tell us the probability the defendant was a contributor. In this case, the LRs of over 1 billion indicate that the DNA evidence provides extremely strong support for the hypothesis that the defendant was a contributor, but this proof is not necessarily definitive and the chances the defendant was NOT a contributor may well be much

higher, or much lower, than one in one billion, depending on what the other evidence in the case might show.

In Europe, where it is more common than in the US for forensic scientists to use LRs to report their findings, experts often use a scale of conclusions to explain the meaning of LRs. A useful example is the scale shown in Table 1, which was proposed by the United Kingdom-based Association of Forensic Science Providers:

Table 1. Recommended Likelihood Ratio Terminology (AFSP, 2009)

Likelihood Ratio	**Verbal Expression (Strength of Support)**
1–10	Weak or limited support
10–100	Moderate support
100–1,000	Moderately strong support
1000–10,000	Strong support
10,000–1,000,000	Very strong support
>1,000,000	Extremely strong support

Scales of this type may help communicate to lay audiences that LRs are statements about the strength of the evidence for ***supporting*** a particular hypothesis, relative to an alternative hypothesis, rather than statements about the likelihood that a particular hypothesis is true.

V. Evaluating the Evidence

The law requires that expert testimony "is the product of reliable principles and methods" (Rule 702(c)), and that "the expert has reliably applied the principles and methods to the facts of the case" (Rule 702(d)). To apply these legal rules, courts must first identify "the principles and methods" that underlie the proferred expert testimony. A threshold issue in this case is what constitutes "the method." Is it probabilistic genotyping in general, or the specific

PG program that was used (in this case STRmix™), or (even more specifically) the particular version of that program that was used (in this case STRmix™ Version 2.4)?

If the goal of the evaluation is to assure the trustworthiness of the evidence presented to the jury, then a focus on specific program used in the case, in this instance STRmix™, appears to be the right level of analysis. The expert testimony in the evidentiary hearing, and the validation studies presented, all concerned STRmix™ specifically. The record in this case does not allow an assessment of other PG systems, such as TrueAllele™. Moreover, the testimony and evidence points strongly to the conclusion that different implementations of PG systems may differ in their accuracy and their range of validity. Evidence proving that one PG program works well is not sufficient to prove the validity of all other PG systems. Each must be validated separately.

Whether the analysis should be limited to a specific version of STRmix™ poses a similar question about the right level of analysis. This question arose during the evidentiary hearing when evidence emerged indicating that the Midwestern Regional Forensic Laboratory conducted all of the analyses in this case using STRmix™ Version 2.4, while the major validation study offered to show the accuracy of STRmix™ (Bright et al., 2018) was based on analysis with STRmix™ Version 2.5. The answer to this question depends on whether the shift from version 2.4 to version 2.5 could have significantly affected the accuracy of STRmix™. Dr. Buckleton testified that STRmix™ is updated annually, and that Version 2.5 was released approximately one year after Version 2.4. He testified that Version 2.5 incorporated changes in the user interface, offered improved internal diagnostics, and improved processing speed, but changed nothing fundamental in the program that would be likely to affect its accuracy. The defense

offered no evidence to challenge this claim. Because version number appears not to be important for assessing the program's accuracy, this report has referred generically to STRmix™, and has not distinguished among versions. In future cases, however, courts should remain open to hearing evidence that changes in the program may have sufficiently affected its trustworthiness to warrant a fresh review of its admissibility.

In discussing the reliability and trustworthiness of STRmix, this report will focus first on what the PCAST report called "foundational validity"—that is, whether STRmix™ is a method that is capable of producing trustworthy results when used to analyze and compare a suspect to complex DNA mixtures like the one in this case. In *Daubert*, the Supreme Court provided a list of factors for courts to consider when evaluating the reliability of the underlying scientific principles and methods. The following section will evaluate STRmix™ with respect to these "Daubert factors." The final section of this report will discuss what the PCAST report called "validity as applied"—that is, whether STRmix™ was applied properly *in this case*. That will require close examination of the laboratory's "internal validation" as well as various diagnostic indicators that allow assessment of whether the program worked properly in this case. It will also involve discussion of what expert should (and should not) be allowed to say about the meaning of the STRmix™ findings in this case.

A. Foundational Validity

1. Has the Method Been Tested?

The evidence presented in the evidentiary hearing established that STRmix™ can be and has been tested for accuracy. In particular, the studies by Bright et al, and Moretti et al. show persuasively that STRmix™ is capable of producing accurate results with extremely low

error rates: STRmix™ not only works, it seems to work extremely well, at least when used in the manner it was used in these studies.

While the defense experts raised legitimate issues about whether the validation research has gone far enough, their concerns about potential errors are, at this stage, somewhat hypothetical. They are concerned that STRmix™ may prove "brittle" such that it breaks down and produces inaccurate results under specific circumstances that have not yet been identified. These concerns certainly deserve some weight. On the other hand, the studies by Bright et al., and Moretti et al. tested the accuracy of STRmix™ when it was used to analyze well over 2000 known-source DNA mixtures. These mixtures varied in number of contributors (from 3-6) and in the proportion of DNA contributed by each, as called for in the PCAST report. When the mixtures were compared with the DNA profiles of thousands of known contributors and millions of non-contributors, STRmix™ was able to distinguish the contributors from non-contributor with a high level of accuracy. Given the scope of these studies, it seems likely that any serious, systematic problems with the program would have been detected. While it is conceivable that undetected problems might still exist, or that problems might occur episodically under highly specific circumstances, the findings suggest that such problems (if they exist at all) could not be very common.

Whether STRmix™ is vulnerable to error if the analyst is mistaken about the number of mixture contributors was raised as an issue during the evidentiary hearing. To address this issue, Bright et al. used STRmix™ to analyze mixtures based on the apparent number of contributors, which was sometimes less than the true number of contributors to the laboratory-prepared mixtures examined in the study. The apparent number of contributors

was "determined blind by the submitting laboratory following their own standard operating procedure" (Bright et al., p. 12) in order to simulate what might happen in an actual case. The LRs produced under the incorrect assumptions were then compared with LRs produced when STRmix™ was told the correct number of mixture contributors. The results showed that "underestimation of the number of contributors tends to either have little effect on the LR or will tend to exclude known contributors" (Bright et al., p. 16). Underestimating the number of contributors did not result in a significant increase in the LRs assigned to non-contributors, which suggests that it does not increase the risk of false incriminations. In a few instances, however, underestimation caused STRmix™ mistakenly to assign exculpatory LRs to known contributors, which suggests that underestimation may increase the risk of false exclusions.

In a second experiment, Bright et al. simulated a situation in which the analyst overestimated by one the number of contributors to a mixture. This error caused STRmix™ to produce lower LRs for true contributors than would have been the case if the number of contributors was estimated correctly, but it did not cause the LRs to point in the wrong direction (toward exclusion). This error did not "markedly alter" the LRs for non-contributors, which suggests it did not greatly affect the risk of false incriminations.

These findings provide solid empirical support for Dr. Buckleton's claim that "STRmix™ is …robust to small discrepancies in the number of contributors" and that "if you get the number of contributors wrong, the likelihood ratio tends to be conservative." (TR p. 100) Prof. Krane urged that these finding be viewed with caution as indicating general tendencies that may not hold for all cases. His concerns about whether the testing of STRmix™ has gone far enough to uncover all weaknesses were thoughtful and worthy of consideration. Nevertheless, the record

in this case indicates that STRmix™ has been tested. The testing was reasonably rigorous, extensive and relevant to the circumstances of the current case. The excellent performance of the program in these tests strongly supports the claim that STRmix™ is capable of producing trustworthy results.

2. Rate of Error

There are several different ways to assess the error rate of a computer program like STRmix™ that produces LRs. One simple but crude approach is to ask how often the system produces LRs that support the incorrect hypothesis. In criminal cases the potential for a false incrimination is a particularly important concern, so one relevant question is how often, in studies involving known-source mixtures, STRmix™ assigns a LR above one to a non-contributor.[20] The answer to this question may be a bit misleading, however, because a certain number of false inclusions are expected to occur by chance due to coincidental similarity among DNA profiles of different individuals. Indeed, the LR assigned to a particular suspect is, in part, an indication of the likelihood of such a coincidental false identification. "In an experiment on 10,000 false contributors we would expect approximately one LR ≥ 10,000, plausibly 10 above 1000 and 100 above 100." (Bright et al., 2018). So even if STRmix™ is flawlessly determining the genotypes of mixture contributors, we should expect false incrimination of some non-contributors, albeit mostly with relatively low LRs. In the validation studies (Bright et al., 2018; Mortetti et al., 2017), there were indeed many instances in which non-contributors were assigned LRs above one. However, most of these non-contributors were

[20] A LR above one indicates that the peaks observed in the mixture are more common if the person in question is a contributor than if an unknown unrelated individual is a contributor. When the person in question is known not to be a contributor, a LR above one supports the incorrect hypothesis.

assigned low LRs (close to one); which means these misleading STRmix™ results were only weakly incriminating.

It may make more sense to ask how often STRmix™ indicates there is "very strong support" for the wrong hypothesis by assigning a LR ≥ 10,000 to a non-contributor. Bright et al. (2018) report that STRmix™ assigned a LR ≥ 10,000 to a non-contributor only 20 times in their study in which STRmix™ was used to compare approximately 20 million non-contributor profiles to 2825 DNA mixtures. The highest LR assigned to a non-contributor was approximately 500,000. These findings suggest that the rate at which STRmix™ produced "very strong support" for inclusion of the wrong person is extremely low.

There are mathematical tests for determining whether, in general, the LRs assigned by a program like STRmix™ are higher or lower than they should be (Taylor, Buckleton & Evett, 2015). In theory, if the LRs are properly calibrated then the average of the LRs assigned to non-contributors in a study like Bright et al. (2018) should be approximately 1.0. If the average of the non-contributor LR's is higher than one, it indicates the LRs assigned by the program are too high, which means that the LRs overstate the value of the DNA evidence. A related test is to check the proportion of non-contributors who are assigned LRs above a particular level. If the program is properly calibrated, no more than 1 in 10 non-contributors should be assigned a LR above 10; no more than 1 in 100 should be assigned a LR above 100; no more than 1 in 1000 a LR above 1000, and so on. Non-contributors should be assigned LRs greater than one only when they happen by chance to have profiles similar to a true contributor, and that kind of coincidental incrimination should occur at a frequency that is roughly the reciprocal of the LR.

Thus, among non-contributors the proportion assigned a given LR should be no higher than 1/LR.

The findings reported by Bright et al. (2018) indicate that STRmix™ passed both of these tests. The average LR for non-contributors was less than one and the number of non-contributors assigned LRs above one was less than *1/x*, where *x* was the value of the LR. These findings allay concerns that the LRs produced by STRmix, when viewed at an aggregate level, overstate the probative value of the DNA evidence.

Professor Krane urged that these findings be viewed with caution given that the data reported reflect averages and may hide the presence of "outliers"—i.e., uncommon cases in which the LRs could be seriously wrong (TR p. 435-436). He argued that the "failure rate" of a system like STRmix™ is likely to vary across cases and may be higher for "difficult samples where there [are] a large and unknown number of contributors" and may "be exacerbated when there is a significant or an increasing risk of allelic drop-out." (TR p. 436). In such cases, the "parameters" used by STRmix™ to assess genotype weights may not apply and, as a result, the computation of LRs could go awry. Krane's concerns seem consistent with those of other scholars who have expressed concerns about the reliability of LRs computed on the basis of complex models that may, in some instances, be based on faulty assumptions (Lund & Iyer, 2017). These concerns should be taken seriously, and may well warrant cautious treatment and extra scrutiny of "difficult cases" (e.g., those with large numbers of contributors, low-level contributors, and a significant risk of allelic drop-out), as discussed below in the section on Validity as Applied.

The evidence presented in the hearing indicates, however, that the error rate of STRmix™ is likely to be quite low in most cases. Large studies in which millions of non-contributor profiles were compared with DNA profiles of thousands of mixed DNA samples showed that STRmix™ very rarely produced strongly incriminating findings against a non-contributor. Statistical analyses suggest that, in the aggregate, the LRs produced by STRmix™ are properly calibrated and do not overstate the value of incriminating evidence. This evidence strongly supports the claim that STRmix™ is "foundationally valid."

3. Peer Review and Publication

A large number of scientific articles concerning STRmix™ have been published. During the evidentiary hearing, the government offered an exhibit (Government Exhibit 3) that listed 47 peer-reviewed articles on DNA mixture interpretation. Most of these articles were related to probabilistic genotyping and mentioned or discussed STRmix™. Key articles that offer a detailed description of the program and discuss how it works include Taylor, Bright & Buckleton (2013), Bright, Taylor, Curran & Buckleton (2013); and Taylor, Bright, Buckleton & Curran (2014). As already mentioned, there are two key articles testing the accuracy of the program for mixtures of four or more individuals, and both were published in respectable peer-reviewed scientific journals (Bright et al., 2018; Moretti et al., 2017).

4. Standards

There was much discussion of standards during the evidentiary hearing. On balance, the testimony suggested that STRmix™ meets (or comes very close to meeting) standards for probabilistic genotyping systems promulgated by the FBI's Scientific Working Group on DNA Analysis Methods (SWGDAM)(Defense Exhibit 1) and IAFS (Defense Exhibit 21), and possibly

also the standards of the British Forensic Science Regulator (Defense Exhibit 22). On the other hand, STRmix™ has not undergone the stringent verification and validation process specified by the IEEE standards for safety critical systems (Defense Exhibit 20).

The standards of SWGDAM, IAFS and the Forensic Science Regulator apply specifically to the use of PG systems in forensic science and were designed by (or with substantial input from) forensic scientists. The IEEE standards apply more broadly to "safety critical systems" in a variety of domains, including aviation and medicine, and were largely designed by engineers and software designers. While it is tempting to conclude that forensic scientists are in a better position to set standards for forensic science than software engineers, it is not clear that this is true when the standards apply to complex software systems. Software engineers may well be better positioned to understand what might go wrong with such a system and how to minimize those risks. Hence, the failure of STRmix™ to meet (or fully meet) the IEEE standards for safety critical systems deserves consideration. On the other hand, as discussed earlier, concerns about the possibility of undetected points of failure in STRMIX™ are, at this juncture, somewhat hypothetical and must be weighed against evidence that the program has worked well in validation studies.

5. General Acceptance

STRmix™ has been widely adopted by forensic DNA laboratories. Over 40 laboratories in the United States and many additional laboratories internationally have purchased licenses to use STRmix™ , making it by far the most frequently adopted program for probabilistic genotyping. This widespread adoption certainly suggests that it is accepted to be reliable by forensic scientists.

There are, however, dissenting voices who have raised concerns about whether STRmix™ is adequately validated. These concerns were raised in the evidentiary hearing by the defense experts, Professors Krane and Heimdahl, and Mr. Adams. They are not the only critics. The defense pointed to a letter published recently in a forensic science journal by four Australian academics that echoes points made by the defense experts in this case. The letter expressed the opinion that "substantially more evidence is needed to establish foundational validity [of STRmix™] across broader settings." (McNevin, Wright, Chiseling & Barash, 2019)(Defense Exhibit 26). While the critics are largely academics, rather than forensic scientists, it appears that the critics are well qualified to comment on the validation of a software program like STRmix™. Indeed some of the critics, such as Professor Krane who has published a book on "bioinformatics" and Professor Heimdahl, who specializes in the evaluation of safety critical software, are undoubtedly more knowledgeable than most forensic scientists about the potential limitations of such systems and how those systems should be evaluated.

On the other hand, it is important to consider the breadth and scope of the dissenting opinions. These critics are not claiming that STRmix™ is unreliable or invalid. They are merely raising concerns about whether validation efforts (which so far have provided strong support for STRmix™) have gone far enough to detect and root out all possible errors that might occur. These concerns about potential undetected error should not be ignored, but need to be weighed against the empirical evidence that STRmix™ worked well for analyzing a broad range of DNA mixtures similar to the ones at issue in this case.

B. Validity As Applied

The factors discussed in the previous section relate to the "foundational validity" of STRmix™. This section analyzes its "validity as applied"—that is, its trustworthiness *as used in this case*.

As discussed earlier, each laboratory that adopts STRmix™ must perform a number of studies to measure the variability of the DNA profiles produced in that laboratory. These parameters are used to "tune" STRmix™ to the particular set of equipment and operating procedures employed by the laboratory, in order to assure that the program will operate properly. The Midwest Regional Laboratory carried out a series of studies that were specified by the SWGDAM guidelines, and by the developers of STRmix™ (see summary in Defense Exhibit 14). The laboratory then sent the data collected in these studies to ESR (the company that markets STRmix™) for statistical analysis and interpretation. ESR returned a report to the laboratory, indicating what parameters should be used.

Professor Krane criticized this "internal validation" procedure, characterizing it as "almost farcical, ridiculous"(TR p. 382). In his view, it is not acceptable for the personnel of a forensic laboratory to hire someone else to help them interpret their validation data and establish parameters. By hiring someone else to do this work, analysts at the Midwest Regional Laboratory missed a crucial learning opportunity that would have given them "a visceral level of understanding" about how STRmix™ works (TR p. 382). Like students who pay someone else to "do the analyses for them and then hand them the data they can use in a lab report" (TR p. 381), these analysts might have gotten the right answer while remaining ignorant of how that answer was obtained or what it means.

This criticism may well raise concerns about the level of analyst training at the laboratory, but it does not speak to the quality of the laboratory's internal validation. It appears that the proper studies were done. No questions were raised about the qualifications of the ESR personnel who analyzed the data. Professor Krane did not suggest that the parameters that ESR established for the use of STRmix™ in this laboratory were wrong or inappropriate. He was primarily concerned about poor pedagogy, not about scientific error in establishing STRmix™ parameters.

Perhaps the most important concern raised by the defense experts was the alleged failure of the Midwest Regional Forensic Laboratory, and forensic laboratories in general, to establish the boundaries of validity for STRmix—that is, the circumstances beyond which the program will not produce trustworthy results. Prof. Krane argued that the existing validation studies have not done enough to test the accuracy of STRmix™ when dealing with "difficult cases" such as cases where the program is used to support the hypothesis that a suspect was a low-level contributor to a complex mixture where there is significant risk of allelic drop-out and the number of contributors is difficult to determine.

This argument is well-founded and deserves to be taken seriously. It raises the question, however, of whether this case is a "difficult case" that may lie near the boundary of validity for STRmix, or whether it is a case that falls more centrally within the category of cases for which the validity of STRmix™ has adequately been established. A close look at the nature of the DNA evidence in this case suggests it is more the latter than the former. Although this case unquestionably involves a complex mixture, and the number of contributors cannot be known with certainty, the defendant's DNA profile is consistent with the genotypes that

STRmix™ assigned to the primary contributor, who was estimated to have contributed more than 50% of the DNA in the mixture.[21] All of the defendant's alleles across 22 loci were detected in the mixture, so this is not a case where the difference between an incriminating or exculpatory finding depends critically on the correctness of parameters (e.g., those related to related to allelic drop-out rates) that might not have been adequately evaluated for the circumstances at hand. The DNA mixtures analyzed in this case appear similar in all relevant respects to the bulk of mixtures that were successfully analyzed in the validation studies by Bright et al. (2018) and Moretti et al. (2017). There is nothing about the evidence to suggest that this was a particularly difficult or challenging case for STRmix™ to analyze.

Another consideration is whether the STRmix™ analyses conducted in this case were carried out properly. STRmix™ includes a number of important internal diagnostic measures that can be used to assess whether the program operated properly in a given case (Russell et al., 2019). These measures are designed to detect instances where something goes wrong with the modeling process, such as a failure of the MCMC process to identify the most likely explanations for the observed data, or instances where the program needs to make implausible assumptions about the parameters in order to "fit" a model to the data (Russell et al, 2019). The internal diagnostics indicate various underlying assumptions and intermediate conclusions that the program reached in the process of determining the weights to assign to the genotypes of possible contributors. A trained analyst can examine the diagnostic measures in order to assess whether the models generated by STRmix™ are plausible and persuasive explanations for the data observed in the electropherograms, as opposed to being implausible or unlikely

[21] In the best fitting model for Item 8-B, for example, the mixture percentage for the primary contributor was 56%.

explanations that rest on unlikely or dubious assumptions. These measures are also useful for identifying instances in which something might be wrong with the operator inputs to the program, such as making an incorrect assumption about the number of contributors, inputting an incorrect reference profile for a known contributor, or entering incorrect values for the parameters used to "tune" the system to the laboratory. The inclusion of these diagnostic measures in the program is an acknowledgement that the modeling process is not always perfect, and that operator errors can also occur. Hence, it is extremely important that these measures be carefully evaluated by analysts who use STRmix™.

In order for these internal diagnostic measures to serve their purpose it is, of course, important that the analysts operating the program have a thorough understanding of how they work and what they mean. One worrisome aspect of the evidence presented during the evidentiary hearing is that the Technical Leader of the DNA unit of the laboratory appeared to have a limited understanding of the STRmix™ internal diagnostics. For example, when questioned by defense counsel about the meaning of the diagnostic measures provided by STRmix™ in this case, Ms. Anne Ciecko did not know how the numbers for "DNA amounts" were calculated (TR p. 164); how "mixture proportion" numbers were calculated (TR p. 165); or how numbers indicating "degradation" were calculated (TR p. 168-69). She did not know what was represented by numbers indicating the program's "acceptance rate" (TR p. 172), "average (log) likelihood" (TR p. 173); "locus efficiency" (TR p 179); or "minimum allowed variance from the mode" (TR p. 181). She said she "would have to look it up" when asked for explanations of the numbers reported for "locus amplification variance" (TR p. 182), "maximum stutter," and "forward stutter" (TR p. 182). She also testified that she did not know the meaning of the

number provided by the program for "drop-in cap," "drop-in frequency," "drop-in parameters," "RWSD," and "ESS thinning" (TR p. 183).

It is certainly possible that an analyst might operate STRmix™ successfully without full knowledge of the meaning of the diagnostic indicators, just as a pilot might successfully fly an airplane without fully understanding the gauges and warning lights in the cockpit. Like the cockpit instruments, however, the diagnostic measures incorporated into STRmix™ are there for a reason. A pilot who ignores or fails to understand the cockpit instruments is asking for trouble, and so is a DNA analyst who ignores or fails to understand the STRmix™ diagnostics. Consequently, the testimony of the laboratory's Technical Leader (who is responsible for training of DNA analysts in the lab) raises concerns about the ability of the analysts in the laboratory to appreciate and detect problems with STRmix™ should problems arise in future cases. [22]

However, it does not appear that the STRmix™ internal diagnostics signaled any problems in this case. All of the diagnostic indicators appeared to have values within norms discussed in the literature (Russell et al., 2019). Neither Dr. Buckleton nor Dr. Krane was concerned about any of the diagnostic indicators reported in this case. While this is an important issue for counsel to consider in future cases, it appears that STRmix™ was applied properly in this case.

[22] In fairness to Ms. Ciecko, it should be noted that some of these matters could be clarified by consulting with reference works, such as the operator's manual for STRmix or published articles (e.g., Russell et al. 2019). It might also be true that the analyst who actually performed the STRmix analysis in this case was better informed on these issues. But the familiarity of the operator with STRmix diagnostics is an important issue for counsel to consider when evaluating expert qualifications and hence seemed worth highlighting in this report.

A final issue concerns the interpretation of the likelihood ratios produced by STRmix™. Because LRs are often misinterpreted, even by experts, this issue will warrant close attention when and if expert testimony about STRmix™ is presented at trial.

During the evidentiary hearing, the following exchange occurred between the Court and Anne Ciecko, the Technical Leader of the DNA unit of the Midwest Regional Forensic Laboratory:

> Q: And the likelihood ratio that STRmix™ will generate is expressed in terms of the chances that this mixture is of this defendant and three random people versus four random people? Is that the statistic it generates?
>
> A: For this case, yes. (TR p. 288, Lines 3-7).

This answer is incorrect and misleading. As discussed above, the LR generated by STRmix™ is not a statement about the chances that the defendant is a contributor to the mixture. It is, rather, a statement about the relative probability of observing the peaks in the electropherogram of the DNA mixture IF the defendant and three random people were contributors, versus the probability for four random people. The difference between the incorrect and correct interpretations is subtle but may be quite important to a criminal defendant. The incorrect interpretation implies that the probability of the defendant being a contributor has been established by the scientific evidence; the correct interpretation implies that the DNA evidence strongly supports the defendant being a contributor without suggesting that the exact probability been established, which makes it easier for jurors to appreciate that they will need to evaluate that probability in light of all the evidence. For a defendant with a strong alibi, this distinction may be crucial.

The laboratory report correctly states the findings as follows: "This mixture is greater than one billion time more likely if it originated from Kenneth Davon Lewis and three unkown unrelated individuals than if it originated from four unknown unrelated individuals." This is a statement about the probability of the evidence under alternative hypotheses about its origin; it is not a statement about the likelihood that the defendant was a contributor. Ms. Ciecko's statement to the contrary during the evidentiary hearing may have been a mere verbal slip, rather than indicating a fundamental misunderstanding. Avoiding such slips during future testimony will be important, however, both as a matter of sound science and fundamental fairness. If an expert misinterprets the results of a scientific method in a way that leads to incorrect expert testimony, then the expert has not applied the method in a valid manner.

Respectfully submitted,

William C. Thompson
Special Master

October 31, 2019

References

Association of Forensic Science Providers (2009). Standards for the Formulation of Evaluative Forensic Science Expert Opinion, *Science & Justice*, 49: 161-172.

Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble (2016). Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. *BMC Genetics,* article number: 125. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7

Bright, J., Richards, R., Kruijver, M., Kelly, H. McGovern, C. et al. (2018). Internal validation of STRmix™--A multi laboratory response to PCAST. *Forensic Science International: Genetics,* 34: 11-14.

Bright, J.-A., Taylor, D., Curran, J. M., & Buckleton, J. S. (2013b). Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics*, 7(2), 296–304. https://doi.org/10.1016/j.fsigen.2012.11.013

Butler, J. (2012). *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press.

de Keijser, J.W., Malsch, M., Luining, E.T., Kranenbarg, M.W., & Lenssen, D. (2016). Differential reporting of mixed DNA profiles and its impact on jurists' evaluation of evidence: An international analysis. *Forensic Science International: Genetics*, 23: 71-82.

Dror, I.E., Thompson, W.C., Meissner, C.A., Kornfield, I., Krane, D., Saks, M. & Risinger, M. (2015). Context management toolbox: A Linear Sequential Unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *Journal of Forensic Science, 60*(4), 1111-1112. Doi:10.1111/1556-4029.12805

Krane D.E., Ford S., Gilder J., Inman K., Jamieson A., Koppl R., et al. (2008). Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. *Journal of Forensic Science, 53*(4),1006–1007. doi.org/10.1111/j.1556-4029.2008.00787.x

Kruschke, J.K. (2011). *Doing Bayesian Data Analysis*. Academic Press.

Lund, S.P. & Iyer, H. (2017). Likelihood ratio as weight of forensic evidence: A closer look. Journal of Research of National Institute of Standards and Technology, 122, 27. https://doi.org/10.6028/jres.122.027

Moretti, T.R., Just, R.S., Kehl, S., Willis, L.E., Buckleton, J.S., Bright, J., Taylor, D.A. & Onorato, A.J. (2017). Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 29:126-144.

Morrison, G.S. & Thompson, W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science & Technology Law Review*, 18: 326-433. http://www.stlr.org/download/volumes/volume18/morrisonThompson.pdf

National Commission on Forensic Science (NCFS). (2015). Subcommittee on Human Factors- *Ensuring that Forensic Analysis is Based Upon Task-Relevant Information*, available at: https://www.justice.gov/ncfs/file/818196/download

President's Council of Advisors on Science and Technology (PCAST) (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*. Executive Office of the President, September 2016. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

President's Council of Advisors on Science and Technology (PCAST) (2017). An addendum to the PCAST report on forensic science in criminal courts. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf

Risinger DM, Saks MJ, Thompson WC, & Rosenthal R (2002). The Daubert ⁄ Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review, 90*, 1–55.

Russell, L., Cooper, S., Wivell, R., Kerr, Z., Taylor, D., Buckleton, J. & Bright, J. (2019). A guid to results and diagnostics with a STRmix™ report. *WIREs Forensic Science*, e 1354. https://doi.org/10.1002/wfs2.1354

Stoel, R.D., Berger, C.E., Kerkhoff,W., Mattijssen, E., & Dror, I. (2015). Minimizing contextual bias in forensic casework. In. Strom K. and Hickman, M.J. (eds) *Forensic Science and the Administration of Justice.* New York: Sage.

Taylor, D., Bright, J. & Buckleton, J. (2013). The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7(5): 516-528.

Taylor, D., Bright, J.-A., Buckleton, J., & Curran, J. (2014). An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. *Forensic Science International: Genetics*, 11, 56–63. https://doi.org/10.1016/j.fsigen.2014.02.003

Taylor, D. Buckleton, J. & Evett, I. (2015). Testing likelihood ratios produced from complex DNA profiles. Forensic Science International: Genetics, 16: 165-171.

Thompson, W.C. (2009). Painting the target around the matching profile: The Texas sharpshooter fallacy in forensic DNA interpretation. *Law, Probability and Risk, 8*, 257-276. doi.org/10.1093/lpr/mgp013

Thompson, W.C. (2018). How should forensic scientists present source conclusions. *Seton Hall Law Review, 48*: 773-813.

Thompson, W.C., & Newman, E. J. (2015). Lay understanding of forensic statistics: valuation of random match probabilities, likelihood ratios, and verbal equivalents. *Law & Human Behavior*, 39(4), 332–349.

Thompson, W.C., Vuille, J., Taroni, F., & Biedermann, A. (2018). After Uniqueness: The Evolution of Forensic Science Opinion. *Judicature,* 102(1): 18-27.